Pitch Extraction Accuracy for Two Voice Analysis Systems

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation
with research distinction in Speech and Hearing Science in the undergraduate
colleges of The Ohio State University

By

Allison Masty

The Ohio State University
June 2007

Project Advisor: Professor Michael D. Trudeau, Department of Speech and Hearing Science

Abstract

In the clinical measurement of voice, the accuracy of the software algorithm used to extract voice parameters is of crucial importance. However, software systems vary in the methods used to extract these parameters, in particular voice fundamental frequency (F0). The present study compared measurements of mean F0 from two such systems, Cool Edit 2000 (now Adobe Audition, Version 2.0) and the Computerized Speech Laboratory, Model 4500 (CSL 4500). Data from 8 trained singers in a pitch-matching task were analyzed by both systems. Surprisingly, results showed that mean F0 estimates calculated by the CSL 4500 were farther from the target frequency of the eliciting stimulus, 207.6 Hz, than were mean F0s calculated by Cool Edit 2000. Neither the type of eliciting stimulus (pure tone or synthesized voice) nor stimulus duration (50, 100, 200, and 300 ms) had an effect on F0 estimates. Results are discussed in terms of user strategies for F0 extraction with these software packages, and in terms of implications for the clinical measurement of voice.

Acknowledgements

First and foremost, I would like to thank my project advisor, Dr. Michael Trudeau, for his patience, kindness, and guidance over the past several months. In challenging me to think for myself, you have truly made the completion of this project an excellent learning experience. Thank you for all of your encouragement and constant support; it has not gone unnoticed.

I would also like to extend my deepest appreciation to my honors advisor, Dr. Janet Weisenberger, for introducing me to the challenge of completing a Senior Honors Thesis. For several years, you have been a true inspiration to me, and I greatly appreciate all of the time, effort, and genuine interest you have invested into both this project and my overall academic success at The Ohio State University. There were moments when I truly could not have done it without you.

In addition, I would like to thank Dr. Trudeau, Dr. Weisenberger, and Dr. Karen Peeler for sitting on my defense committee. I greatly appreciate the time you have taken to be a part of my project.

I would like to extend a special thank you to both Stephanie Finical and Stephanie Joseph, my undergraduate colleagues in the Department of Speech and Hearing Science. Thank you for all of the hours you put into data collection, analysis, and organization for this project. Working with you has made my thesis experience especially enjoyable, and I am so glad that we have had the chance to get to know one another over the past several months. I wish you the best of luck in your future endeavors.

Finally, I would like to thank my parents, Jerry and Patricia Masty, in addition to each of my five sisters for always believing in my gifts and talents. Your confidence in my abilities has always been the motivation for my success, and your love and support have made me who I am today.

Table of Contents

Abstract……ii

Acknowledgements……..……..iii

Chapter 1: Introduction and Literature Review…...……...1

Chapter 2: Method……..……..10

Chapter 3: Results……16

Chapter 4: Discussion and Future Research Implications……20

Chapter 5: Conclusion……...25

References……..26

Chapter 1

Introduction and Literature Review

The ability to match pitch to another tone is a highly valued skill for trained singers. In essence, matching pitch is the quality of singing on key, a skill routinely practiced by trained singers. Several previous studies have examined the accuracy of this ability, referred to as pitch-matching accuracy (PMA), as a function of stimulus type, duration, and frequency, among other factors. Fewer studies have examined the means used to extract and measure the fundamental frequency of a signal, which is the basis for clinically judging PMA.

The ability to complete a pitch-matching task requires one to hear and discriminate the frequency of the stimulus to which the subject is matching pitch, and then vocally produce a sound of the same perceived pitch, or frequency. Thus, a pitch-matching task is one of both sound perception and vocal production (Moore, Keaton, & Watts, 2006). In studying the dynamics of accurately matching the pitch of a stimulus token, there are several factors that may influence the outcome. Some of these factors have already been investigated, but it has yet to be determined precisely which factors have greatest influence on measures of PMA.

One such factor investigated in the present study is the accuracy of the tool used to calculate the average fundamental frequency of a signal. In past studies, responses to stimulus tokens of a specific average frequency produced by subjects were analyzed for their accuracy in matching the stimulus frequency across a specified number of cycles in the response token. However, only in Curran's 2004 study has the entire response token been used to analyze the mean fundamental frequency of the response in PMA tasks. She

found, using the Cool Edit 2000 software (which has since been incorporated into the up-and-coming voice analysis system Adobe Audition,Version 2.0), that analyzing the fundamental frequency of the entire response produces the most accurate PMA results as compared to analyzing the fundamental frequency at either the response onset or across the first five cycles of the response (Curran, 2004). The present study measured responses in the data set collected in a study conducted by Ives in 2002 and also used in Curran's 2004 study, using more expensive, more precise, and questionably "more sophisticated" software: The Computerized Speech Laboratory, Model 4500 by Kay Elemetrics.

Factors Influencing PMA

A study of considerable importance in the field of PMA was conducted by Thomas Murry (1990). In this study, Murry investigated the PMA abilities of both trained singers and untrained singers, measuring laryngeal control during the subjects' vocal productions as a function of their experience with singing and natural talent. Presenting the subjects with stimuli of three separate pure tone frequencies of the same duration, Murry instructed subjects to sing the same pitch they heard. As suspected, the trained singers matched the pitch of the stimulus token more accurately than the untrained singers did, suggesting that they possess greater laryngeal control while producing sound than the untrained singers. More relevant to the discussion of the present project, Murry also found that estimates of the singers' pitch-matching abilities measured by the C-Speech vocal analysis software program were more accurate when the average fundamental frequency for the first five cycles of the subject's vocal response

was used for PMA analysis compared to when only the first cycle of vocal production was used to analyze PMA. However, this result proved statistically insignificant.

Although statistically insignificant, the trend towards greater accuracy as more cycles of the response are evaluated raises questions. What is the cause of this slightly more accurate response when PMA is measured across the first five cycles of response instead of just the first cycle of response? Is there truly a difference in phonatory accuracy as the response progresses, or are fundamental frequency measurement algorithms influenced by the increase in available information?

Related to this issue is the question of whether the duration of the eliciting stimulus affects response accuracy. Tervaniemi et al. (2000) investigated this question; his results suggested that durations of the stimulus token exceeding 100 ms have little effect on subjects' pitch discrimination abilities. This study also found that pure tone stimuli make poorer stimulus tokens than do more spectrally rich stimulus tokens for evoking an electrophysiological pitch discrimination response. This finding is important because it suggests that subjects would be better able to accurately complete a pitch-matching task when the stimulus is something more complex than a pure tone, for example music or speech.

Ives further investigated this question of the effect of stimulus duration, as well as token type, on the production aspect of PMA in his 2002 study. Ives presented subjects with stimulus pure tone and synthesized voice tokens of various durations, hoping to discover a link between PMA and token type and duration. In fact, Ives found that token type did not significantly affect PMA, nor was there a significant effect of token duration on PMA. However, a significant difference was observed between PMA measured

across the first cycle of subject response as compared to the first five cycles of subject response. This implies that PMA is not affected by stimulus token durations that exceed 50 milliseconds, but that the number of cycles in the response duration for which the fundamental frequency is measured is a significant factor in PMA. When measured across the average fundamental frequencies of the first five cycles, Ives found that PMA was much closer to the stimulus frequency than PMA measured across the first cycle. One interpretation of this finding is that a period of stabilization exists near the beginning of the subjects' responses during which the subject vocally "finds" the desired frequency and produces it.

Ameer (2003) re-analyzed Ives' data and confirmed the existence of a stabilization period in PMA tasks during which subjects appear to adjust their response to best match the pitch of the stimulus token. Interestingly, Ameer found that this stabilization period spans the first six cycles of subject response, no matter what the stimulus type or duration was. The implications of this finding support Ives' assertion that stimulus token duration does not affect PMA. Ameer's results do suggest that measuring the mean fundamental frequency only through the fifth cycle of response to calculate PMA for the entire response may lead to inaccurate conclusions.

As noted above, Curran (2004) analyzed PMA across three lengths of subject response: the first cycle, the first five cycles, and the entire subject response and found that measurement of the entire subject response produces the highest accuracy in PMA. However, in contrast to Ives' (2002) results, Curran also found that 50 ms stimulus tokens were only of sufficient duration to produce accurate PMA if the token was spectrally rich (Curran, 2004). In instances where pure tones were the presented

stimulus, greater durations resulted in more accurate mean fundamental frequencies of subject responses.

In conducting their studies, Curran (2004) and Ives (2002) used the software Cool Edit 2000. Curran noted a potential problem with the Cool Edit pitch extraction algorithm. Cool Edit 2000 analyzes fundamental frequency based on zero crossings, which in some cases yielded inaccurate results for the mean fundamental frequency of subject responses that had more than one zero crossing. In effect, this algorithm analyzed the fundamental frequency of such complex subject responses as two separate waveforms instead of one complex waveform, almost doubling the extracted fundamental frequency of the response. To correct for this, Curran divided the computed frequency for these cases by two. However, this procedure may not have produced an accurate result.

Methods for Estimating Voice Pitch

Curran's observation of potential problems in the Cool Edit pitch extraction subroutine points to the importance of evaluating the tools used to measure the mean fundamental frequency of a response in a pitch-matching task, namely, the computer software used to analyze the produced signal. Because physically counting the cycles in a specific acoustic waveform can be tedious and at times difficult due to the complexities of the human voice, mean fundamental frequency analysis has turned to the use of technology. There are several commercial software programs available for analyzing an acoustic signal, and many clinicians rely on these tools for accurately evaluating an acoustic signal. Many of these programs employ one or more pitch extraction algorithms for calculating fundamental frequency (Read, Buder, & Kent, 1992). These algorithms calculate fundamental frequency in slightly different manners, thus suggesting that

fundamental frequency extraction is a fluid exercise with many different techniques that, theoretically, produce the same desired result.

Upon further investigation, it becomes clear that the process that the computer goes through when a clinician inputs a signal into a program is more complicated than the few seconds it takes for the program to compute mean fundamental frequency would suggest. In fact, the most widely used fundamental frequency extraction algorithms belong to one of two groups: those that use event detection methods and those that use short-term average methods. In event detection method algorithms, fundamental frequency calculations are based on concrete events, like positive or negative waveform peaks, or zero-crossings of the waveform. Contrastingly, short-term average methods require the computation of fundamental frequency over a "short sliding window of the input data" (Parsa & Jamieson, 1999). The data that are input in the short-term average methods may be information relating to either time or frequency, and the small window allows pitch contours to remain virtually identical to the true pitch contour being measured (Parsa & Jamieson, 1999).

In a 1999 study by Parsa and Jamieson, seven high precision pitch extraction algorithms were tested under various background noise and vocal perturbation conditions to evaluate their performance in calculating mean fundamental frequency. In clinical practice, period-to-period variations in the fundamental frequency of a vocal production are usually indicative of a vocal pathology, thus making highly accurate fundamental frequency estimation extremely important (Parsa & Jamieson, 1999). Parsa and Jamieson found that time domain techniques are better able to produce accurate fundamental frequency estimates than estimates relying on the information in the frequency domain.

However, Parsa and Jamieson's results also show that frequency domain techniques are better able to handle changes in the amplitude of the input waveform. These results suggest that the importance of accuracy in both the time and frequency domains cannot be overlooked.

Read, Buder, and Kent (1992) also explored the performance of various fundamental frequency algorithms, but through comparing the performance of seven different systems marketed for acoustic analysis in several different analysis areas. Overall, the investigators' findings suggest that systems should provide at least two separate methods of fundamental frequency extraction with at least one of the time-domain type and the other of the short-term average type (Read, Buder, & Kent, 1992). The investigators further state that fundamental frequency extraction is one of the most concerning functions of the various systems, for subtle errors in this area are difficult to detect, even to the well-trained clinician.

Interestingly, both the Parsa and Jamieson (1999) and the Read, Buder and Kent (1992) studies point out a flagrant problem with most acoustic analysis systems: system developers rarely publish or document the fundamental frequency algorithms employed to analyze data. Such practices make choosing an acoustic analysis program to meet specific needs most difficult, especially given that different algorithms perform differently under different circumstances. Thus, choosing the right tool for the task is nearly impossible. In addition, the cost of acoustic analysis programs varies widely from a free software download to thousands of dollars for a console plus software. It is becoming increasingly necessary for consumers to know the specifications of the acoustic analysis system they are purchasing.

One of the systems tested in the Read, Buder, and Kent, study (1992) was the Kay 5500. The investigators then suggest that users of the Kay 5500 may be interested in trying the then recently released Computerized Speech Laboratory (CSL) from the same company (Read, Buder, & Kent, 1992). For the present study, a version of the CSL, Kay Elemetrics' CSL, Model 4500 (CSL 4500), was used to analyze the pitch-matching response tokens collected in Ives' 2002 study. One feature of the CSL 4500 that is particularly relevant to the present study is the pitch extraction options. Users of the CSL 4500 may choose whether to use event detection or short-term average methods in mean fundamental frequency analysis, which adds an important level of sophistication to the CSL (A. Bohman, personal communication, May 17, 2007). Because this version of the CSL is considered to be one of the most sophisticated speech analysis systems in the industry, it is expected to produce precise and accurate estimations of mean fundamental frequency. Such a finding may suggest that extremely sophisticated instrumentation is a necessary part of increasing the accuracy of clinical practice, perhaps improving the ability of clinicians to help patients suffering from various vocal pathologies.

Like Curran's study, the present project investigated the question of PMA as a function of the mean fundamental frequency across the entire subject response to stimulus tokens of various types, durations, and frequencies. However, the above-noted limitations with the Cool Edit 2000 software underscore the need to confirm the findings from Curran's 2004 study by using more advanced instrumentation. Data from Ives' study (2002) were used to analyze the PMA of trained singers presented with different stimulus tokens (pure tone and synthesized voice) over a range of stimulus frequencies and durations. In contrast to Curran's use of Cool Edit 2000, the present study used the

Computerized Speech Laboratory, Model 4500 software by Kay Elemetrics to analyze the aforementioned data. The hypothesis is that PMA as measured over the entire response using the CSL 4500 will be more accurate than PMA as measured with Cool Edit 2000. This accuracy will be reflected in estimates of the mean fundamental frequency of responses that are close to the frequency of the eliciting stimulus.

Chapter 2

Method

This study analyzed data created and collected by Shawn Ives for use in his 2002 study of stimulus sound duration and spectral complexity as they relate to pitch matching accuracy. The present study compares measurements of specific associated features of pitch matching accuracy (i.e. response duration and mean fundamental frequency of response) obtained by Ives (2002) using the Cool Edit 2000 computer program with newly obtained measurements of these features using Kay Elemetrics' Computer Speech Laboratory, Model 4500. All information for the present study regarding test subjects, token recording procedures, and the task performed by subjects was gathered from Ives' study (2002).

Subjects

Ten male college students, at either the graduate or undergraduate level, with a vocal music background were selected as test subjects for this study. No subject had less than four years of previous formal vocal instruction or less than six years of choral singing experience. All subjects had suspected normal vocal function with no reported history of laryngeal pathology or voice disorders requiring phonosurgery or voice therapy. These steps established that all subjects were trained singers with "normal" vocal production abilities, and a hearing screening was then conducted on each subject to establish that his sound detection abilities were at a "normal" level. Indeed, no subject had a hearing threshold greater than 20 dB HL at all tested frequencies. Subjects were then instructed to deliberately warm-up their voices for no less then 10 minutes at least two hours before participating in the experimental task. Upon completion of the vocal

exercises, subjects were seated comfortably in a sound-treated room and instructed to listen to the tones as they were presented. In response to each presented tone, subjects were to produce a vocal response of the same pitch as the stimulus token as quickly and as accurately as possible.

At this time, each subject was familiarized with the experimental task through a training period of five trial tokens. Each of the five trials ranged in duration from 50 to 300 milliseconds, and in pitch from 130.01 to 311.1 Hz by random selection. None of the five trial tokens was of the same fundamental frequency as the that of the stimulus tokens to be used in the actual recorded experimental task, thus controlling for pitch predictability based on the trials presented during the training period. Subjects were informed that they could request an amplitude adjustment of stimulus tokens at any time during the recording sessions, but no subject did so.

Creation of Stimulus Tokens

All stimulus tokens were one of two types: pure tone or synthesized human voice. There were eight target frequencies between 130.1 Hz and 311.1 Hz presented to the subjects in separate tokens, each frequency commonly perceived as one of the following pitches: C, D, E, F#, G#, A#, C#, and D#. The formants and bandwidths used to create the synthesized voice tokens were established from a vocal sample provided by a professional Bass who is a faculty member of The Ohio State University. Based on this sample, 554 Hz, 916 Hz, and 2466 Hz were established as the first three formants of synthesized voice samples. In addition, harmonics of the fundamental frequency of the token up to 3400 Hz were represented, as well as a 10 cycles-per-second vibrato rate to further simulate the male singing voice in the synthesized tokens. This vibrato rate was

created by shifting the fundamental frequency of the token by half a semitone above and below the established fundamental frequency. All respective pure tone tokens were directly generated by Cool Edit 2000.

There were four token durations: 50, 100, 200, and 300 milliseconds. These durations were chosen based on limits established in other studies with 50 milliseconds being the briefest duration acceptable for a pitch-matching task without resulting in an audible click, and 300 milliseconds being a sufficient representation of a long duration in pitch matching tasks (Tervaniemi et al., 2000).

Once token types, frequencies, and durations had been established, a computer-generated random digits list was used to create a randomized sequence of five token sets for each subject to respond to. Random digits lists were also used within the token sets to create randomized sequences of 64 tokens within each set. This ensured that results of the study would not be influenced by effects of the order of stimuli presentation. It also ensured that subjects would be less able to establish a sense of tonal key, which has the potential to influence their pitch-matching ability. Finally, each token set consisted of stimulus tokens representing both token types, all four token durations, and all eight token frequencies, resulting in 320 total stimulus tokens to which each subject responded.

Recording Procedures

Upon completion of vocal warm-ups and after being seated in a sound-treated room, subjects responded to a series of five stimulus tokens marked for a training period, as previously discussed. Subjects were seated at an average distance of 30 to 35 centimeters from a microphone used to record their responses, and presented stimulus tokens via speaker. Subjects were instructed to respond to the stimulus tokens no less

than two seconds after their presentation with responses no greater than two seconds in duration so that their responses remained within the five-second interval between tokens in each set. Subjects were also permitted to produce responses in either a "straight tone" or vibrato fashion at whatever volume they desired, since neither quality of voice nor amplitude of voice were being measured. Furthermore, two windows were opened in Cool Edit 2000. The first window was used to play the sound file for the subject. The second window was utilized as a means to record both the stimulus token and the subjects' responses as a single sound file. Ives (2002) made use of a custom made mixer/splitter to allow for the stimulus sound tokens and subject pitch matching attempts to be mixed and recorded at the same time. This was done at a 44.1 kHz sampling rate with a 16-bit resolution.

Analysis Procedures

For the current study, data samples were used from the responses of eight subjects employed in Ives' 2002 study. Rather than use the Cool Edit 2000 computer program, the present study employed Kay Elemetrics' CSL 4500 to obtain response durations and mean fundamental frequencies of response tokens. Before opening the .wav sound files containing both the stimulus token and subject response created by Ives, the default settings of the program were adjusted for data analysis. First, the pitch extraction algorithm was set to estimate mean fundamental frequency on the basis of zero crossings under the *Options* window. Settings for voiced period marks were changed so that the impulse locations were at zero crossings and the analysis range was changed from the default setting to display a minimum of 70 Hz and a maximum of 500 Hz. The settings for pitch contours were also adjusted as follows: the analysis range was changed to

match that of the voiced period marks, which was 70 to 500 Hz; frame length was changed to five milliseconds, as was the frame advance setting; and the display range was changed to reflect a 0 Hz minimum and a 500 Hz maximum. The "draw dot contour" option was also selected.

After opening a sound file containing one of Ives' 64 stimulus-token-and-subject-response sets in window A of the CSL 4500, the cursors were used to demarcate the start and end of each subject response, one by one. Using the program's *Zoom--view selection* function, the subject's response was then magnified, and the cursors were again set at what appeared to be the onset and endpoint of the response signal. The cursors were reset, and the *view selection* function was utilized again and again until the data was magnified as greatly as possible. At this point, it was assumed that the time stamp at the response onset and endpoint could be most accurately viewed, and response duration was calculated by subtracting the onset time from the endpoint time to the nearest thousandth of a millisecond. Voiced period marks were superimposed on the waveform sample, but it was common for voiced period marks to appear on the screen only for the first, and part of the second, responses in a file. To further analyze the selected data, a dot contour representing pitch was opened in Window B. By then using the *Statistics* icon in the CSL 4500 toolbar, the dot contour was analyzed for the mean fundamental frequency of the selected signal. On a second computer, a spreadsheet using Microsoft Excel was created in which to record data obtained from analyzing the subject responses, including mean fundamental frequency, onset time, endpoint time, duration, and whether or not voiced period marks appeared in Window A.

Occasionally, the subject response was of such low amplitude that pitch could not even be tracked using the five millisecond settings for frame length and frame advance. In these cases, it was necessary to reset the framing settings so that voiced period marks could be used to calculate the fundamental frequency of the response. Then, the sound file could be opened again in Window A with visible voiced period marks superimposed on the response waveform in question. For a few trials which failed to produce a pitch contour when both frame length and frame advance were set at five milliseconds, the frame length was set at 25 milliseconds and the frame advance at 20 milliseconds. However, when this still failed to produce a pitch contour in window B for a few trials, it was determined by the CSL 4500 user that the voiced period marks should be the definitive framing parameter used to draw a pitch contour when the initial five millisecond frame length and five millisecond frame advance settings failed. Those trials in which the pitch contour for the response had been produced using the 25 millisecond frame length and 20 millisecond frame advance settings were re-analyzed using the voiced period marks framing setting. Thus, in any instance in which the five millisecond default framing failed to produce a pitch contour, the sound file was opened again in Window A, and voiced period marks were used as the framing parameter. In these cases, a note was made in the data spreadsheet that such framing adjustments had been made.

Chapter 3

Results

Mean fundamental frequencies were averaged across all trials of all test subjects for each token type, stimulus token duration, and stimulus token frequency. Average mean fundamental frequencies extracted by Cool Edit 2000 were taken from the results of Curran (2004). To make the present study's data comparable to Curran's, comparisons were only made between pitch estimates extracted using Cool Edit 2000 and those using the CSL 4500 for responses elicited by the 207.6 Hz stimulus token frequency.

Average mean fundamental frequencies for each stimulus token frequency are shown in Table 1. Table 1 also shows the difference in Hz between the estimated mean fundamental frequency and stimulus frequency at each stimulus token frequency.

Token F0	CSL 4500 Mean F0	Difference From Target F0
130.8	275.44	144.64
146.8	291.11	144.31
164.8	299.9	135.1
185	307.67	122.67
207.6	254.58	46.98
233.1	247.95	14.85
277.1	291.03	13.93
311.1	308.22	-2.88

Table 1: Average mean fundamental frequency (F0) and difference from the eliciting stimulus frequency in Hz calculated by the CSL 4500 elicited by each stimulus token frequency across all subjects, trials, durations, and token types.

For every stimulus token frequency except 311.1 Hz, the CSL 4500 estimated mean fundamental frequency of the subjects' responses to be higher than the target frequency. The differences between the target frequency and the pitch estimate decrease

as stimulus token fundamental frequency increases. There is no apparent relationship among the CSL 4500 mean fundamental frequency pitch estimates. For the lower stimulus token frequencies, the difference is considerably larger than it is for the higher stimulus token frequencies.

When comparing the overall mean fundamental frequency of subject response to the 207.6 Hz stimulus token frequency extracted by Cool Edit 2000 and the CSL 4500, the mean fundamental frequency estimated by the CSL 4500 is higher in frequency than that estimated by Cool Edit 2000. Table 2 displays the results of this comparison.

Voice Analysis System	**Mean F0**	**Difference From Target F0**
Cool Edit 2000	214	6.4
CSL 4500	254.61	47.01

Table 2: Average mean fundamental frequencies and differences from the eliciting stimulus frequencies in Hz of the subjects' responses to the 207.6 Hz eliciting stimulus calculated using both voice analysis systems. Values were averaged across all subjects, trials, durations, and token types.

Not only is the mean fundamental frequency extracted by the CSL 4500 higher than the mean fundamental frequency extracted by Cool Edit 2000, but the absolute value of the difference of the pitch estimate is also larger than the absolute value of the difference of the Cool Edit 2000 mean fundamental frequency estimate at 207.6 Hz. However, both the CSL 4500 and Cool Edit 2000 mean fundamental frequency estimates are higher than the eliciting stimulus of 207.6 Hz.

When considering the effects of stimulus token type (either pure tone or synthesized voice) on the pitch estimates extracted by Cool Edit 2000 and the CSL 4500, estimated mean fundamental frequencies are similar within each system. As Table 3

shows below, mean fundamental frequencies calculated using Cool Edit 2000 are still lower than those calculated using the CSL 4500 for both pure tone and synthesized voice stimulus tokens. However, Table 3 also shows that the difference between the pure tone and synthesized voice fundamental frequency estimates are relatively small compared to the differences between the mean fundamental frequency of the subject response and eliciting stimulus displayed in Table 2.

Token Type	**Cool Edit 2000 Mean F0**	**CSL 4500 Mean F0**
Pure Tone	214.8	253.38
Synthesized Voice	213.2	255.84
Difference Between Token Types	1.6	2.46

Table 3: Average mean fundamental frequencies and differences between pure tone and synthesized voice stimulus tokens in Hz for both voice analysis systems elicited by the 207.6 Hz stimulus token. Values were averaged across all subjects, trials, and durations.

Similar to the results in Table 3, Table 4 shows little variation between the mean fundamental frequency estimates extracted by Cool Edit 2000 and the CSL 4500 within each system. However, instead of token type, Table 4 shows the effects of stimulus token duration in ms on the average mean fundamental frequency of the response elicited by the 207.6 Hz tokens estimated by each voice analysis system.

Token Duration	**Cool Edit 2000**	**CSL 4500**
50 ms	214.97	255.74
100 ms	214.81	251.85
200 ms	212.86	254.88
300 ms	213.3	255.85

Table 4: Average mean fundamental frequencies in Hz extracted by both voice analysis systems elicited by the 207.6 Hz stimulus token of the durations shown in ms. Values were averaged across all subjects, trials and token types.

As made evident by the data in Table 4, the CSL 4500 calculated values of mean fundamental frequency to be higher than those calculated by Cool Edit 2000 at each stimulus duration. Both the CSL 4500 and Cool Edit 2000 mean fundamental frequency estimates are higher than the 207.6 Hz target stimulus frequency. Even so, the range of values between the highest and lowest average mean fundamental frequencies within each system is not greater than 2.11 Hz for Cool Edit 2000 estimates, nor 4.00 Hz for CSL 4500 estimates.

Chapter 4

Discussion and Future Research Implications

Because the CSL 4500 is such a widely used, highly reputable, and expensive voice analysis system, it is surprising that Cool Edit 2000, a free software download at the time of Curran's study (2004), produced mean fundamental frequency estimates that were closer to the stimulus token frequency than the CSL 4500 estimates were. Given that the vocal responses analyzed for mean fundamental frequency were produced by trained singers with a relatively large singing history, it is unlikely that the singers' responses were actually as far from the target frequencies as the CSL 4500 mean fundamental frequency results suggest. This raises questions about the results of the study, inviting the consideration of several factors that may have played into yielding these results.

One explanation for the more accurate measures of mean fundamental frequency produced by Cool Edit 2000 is that the mean fundamental frequency extraction algorithm used by the Cool Edit 2000 software to extract pitch may be more precise than that of the CSL 4500. Although possible, this is an arguable explanation. The CSL 4500 is a sophisticated tool heavily relied upon by many clinicians across the country, and it is one of many new editions of the first Computerized Speech Laboratory, suggesting that many of the software's quirks have been revised to near-perfect standards. The CSL 4500 also measures information in the time domain to several decimal places past the thousandths position when measuring in ms, leading one to assume that the CSL 4500's precision of duration measurements is extremely high. Because the frequency of a waveform is

calculated by dividing 1 by the period (1/T=F0), accurate time measurements are imperative for calculating accurate frequency estimates of the waveform.

One would expect that this precision, as made evident in the software's capability to measure duration, would extend to the intricacies of the algorithm or algorithms used to extract pitch by the CSL 4500. However, such a hypothesis is extremely difficult to test due to the proprietary nature of the algorithms themselves. Many voice analysis software companies do not publish the type of algorithms used in pitch extraction, let alone the specific algorithm employed, making the direct comparison of two pitch extraction algorithms employed by voice analysis systems virtually impossible. Such a problem creates a study in which comparing just the results of software data analysis is the objective, leaving little concrete evidence to explain the reasons for differences in pitch extraction results.

Given that the various algorithms employed in fundamental frequency extraction tasks have differing strengths and weaknesses, it follows that one algorithm may be better suited to analyze a sample in a specific situation than another. Thus, in clinical applications, knowing the actual algorithm used in fundamental frequency extraction by a specific software would be beneficial to most accurately assess vocal productions. Future research could focus on obtaining the fundamental frequency algorithm employed by the CSL 4500 and comparing results obtained by its use with results obtained using a different voice analysis system that calculates fundamental frequency based on a different algorithm. If not with a different voice analysis system, the fundamental frequency of the responses could be recalculated using the CSL 4500's short-term average pitch extraction

algorithm instead of the zero crossings algorithm, and the accuracy of results could be compared.

Although it would have been helpful to know the specific algorithm used to calculate fundamental frequency by the CSL 4500 in the present study, the type of algorithm employed in pitch extraction may have also contributed to the great variability in mean fundamental frequency estimates for each target frequency. The CSL 4500 was set by its users in this study to estimate mean fundamental frequency on the basis of zero crossings, an event detection method for pitch extraction, to keep the resulting data comparable to Curran's (2004) mean fundamental frequency values estimated by the Cool Edit 2000 software. It is likely that a second explanation for the disparity in results between the CSL 4500 and Cool Edit 2000 arose from not properly adjusting the data to account for specific incompatibilities between the two systems. All of the data used in the present study were recorded using Cool Edit 2000 by Ives (2002). Ives was able to amplify the recorded signal as part of his pitch extraction procedure, thereby creating a waveform sample of each subject response with a characteristic amplitude that was high enough to visibly count the periods of the waveform. Such amplification was impossible with the CSL 4500. The visual sample of each subject response was barely visible once the file was opened the CSL 4500 software, and the individual periods of the waveforms could not be seen, even using the CSL 4500's *Zoom* function. Therefore, amplifying the signal before analysis using the CSL 4500 may better preserve the integrity of the signal and eliminate this problem.

Because the amplitudes of the sample waveforms were so low when the sound files were opened using the CSL 4500, it is possible that the signal was not strong enough

for the CSL 4500 to accurately track and measure the pitch of the response signal. Indeed, while conducting the present study, there were many instances in which it was impossible to extract the mean fundamental frequency of a sample using the zero crossings algorithm and default 5 millisecond framing window and 5 millisecond frame advance settings. In order to obtain a frequency reading in these instances, it was necessary to change the settings, thus creating inconsistencies in the methods used to obtain each of the mean fundamental frequency estimates.

To further complicate the issue, there exists no standardized set of instructions for handling situations in which the CSL 4500 cannot calculate fundamental frequency. Thus, judges using the CSL 4500 in the present study adjusted for this software limitation in different ways, either by changing the framing window and advance settings, using voiced period marks instead of the pitch contour, or leaving data entries blank in pitch estimations. These differences between users created a data set developed through inconsistent methodologies. Additionally, judges of the subject response durations in the present study did not know the target frequency used to elicit the response being measured. This resulted in the incapability of the judges to know when the mean fundamental frequency estimates calculated by the CSL 4500 were extremely different from the stimulus token fundamental frequencies.

Not only were these mean fundamental frequencies far from the target stimulus frequencies, but data from Joseph (2007) confirm that the estimated mean fundamental frequencies were also highly variable for samples analyzed by more than one user of the CSL 4500. Joseph examined the interjudge reliability of results obtained by three users of the CSL 4500. She found that the three judges' results were highly consistent in

measures of duration, but extremely variable in estimations of mean fundamental frequency. One can conclude that this variability in mean fundamental frequency estimates is the result of using different settings for the analysis of the mean fundamental frequencies of responses. This is an issue that did not arise in Curran's (2004) or Ives' (2002) studies, since each employed only one judge in the task of measuring the duration and mean fundamental frequency of each subject response.

Future studies could handle these problems by focusing on the standardization of instructions for CSL 4500 users so that the effects of differing interjudge methodologies could be ruled out. Since the default settings of the CSL 4500 did not produce pitch contours, voiced period marks, or fundamental frequency results for some waveforms, standard instructions for managing difficult-to-detect waveforms would be beneficial. Moreover, future studies could also attempt to amplify the signal before measuring its fundamental frequency using the CSL 4500. By pre-amplifying the signal, the individual peaks, valleys, and zero-crossings of each the sample waveforms will likely become more visible in the CSL 4500. With increased visibility of the distinct events in the waveform, the characteristics of the individual waveform may be better detected by the software. This improvement in waveform characteristic detection may produce mean fundamental frequency estimates that are more accurate to the true fundamental frequency of the vocal response produced by the trained singer in a pitch-matching task.

Chapter 5

Conclusion

Because mean fundamental frequency estimates extracted using the CSL 4500 varied so greatly from those extracted using Cool Edit 2000, the results of the present study suggest the need for a better understanding of voice analysis systems. The selection of voice analysis software can yield highly variable results in the estimation of mean fundamental frequency, making user familiarity and expertise with the complexities of a specific software of paramount importance in obtaining accurate results. Although the CSL 4500 is a sophisticated tool for analyzing various parameters of vocal productions, the sensitivities of the software to input data, user methodology, and settings makes it a difficult system for the novice user. Indeed, when employed in clinical voice settings, the results of the present study suggest that great caution should be exercised when using the CSL 4500 to extract pitch estimates.

References

Ameer J. Time to phonatory stability in trained singers cued for pitch matching by pure-tone and synthesized voice. Unpublished Honors Thesis, (2003) The Ohio State University.

Curran K. Measurement of Pitch Matching Accuracy: How Much Is Too Little? Unpublished Honors Thesis, (2004) The Ohio State University.

Ives S. The Effects of Sound Duration and Spectral Complexity on Pitch-Matching Accuracy in Singers When Cued with Pure-Tone and Synthesized Human Voice Models. Unpublished Master's Thesis, (2002) The Ohio State University.

Joseph S. Interjudge Reliability in the Measurement of Pitch Matching. Unpublished Honors Thesis, (2007) The Ohio State University.

Moore R, Keaton C, Watts C. The Role of Pitch Memory in Pitch Discrimination and Pitch Matching. *Journal of Voice*. 2006 May 25; [Epub ahead of print].

Murry T. Pitch-matching accuracy in singers and nonsingers. *J Voice*. 1990; 4 (4): 317-21.

Parsa V, Jamieson, D. A comparison of High Precision F0 Extraction Algorithms for Sustained Vowels. *Journal of Speech, Language, and Hearing Research.* 1999; 42: 112-126.

Read C, Buder EH, Kent RD. Speech analysis systems: an evaluation. *Journal of Speech and Hearing Research.* 1992; 35 (2): 314-32.

Tervaniemi M, Schroger E, Saher M, Naatanen R. Effects of spectral complexity and sound duration on automatic complex-sound pitch processing in humans- a mismatch negativity study. *Neuroscience Letters*. 2000; 290: 66-70.